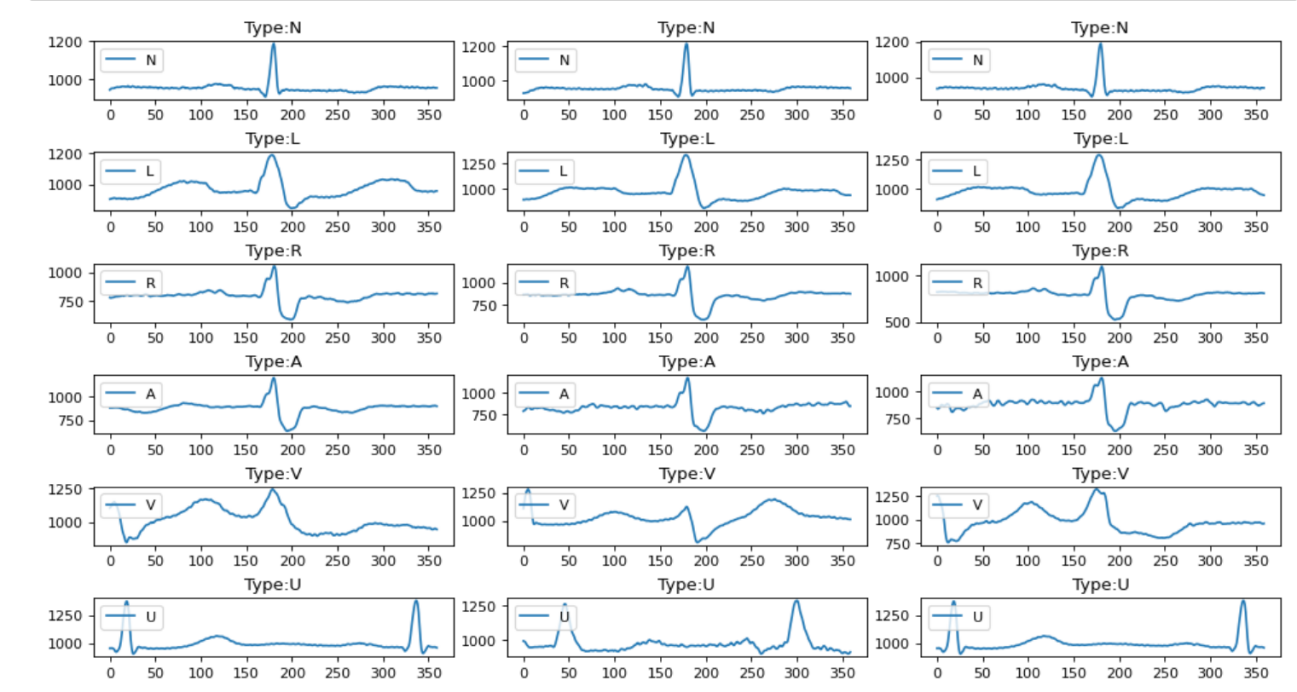
Mini Project1 Report

Xiangyu Gao & Rui Li

**Answer1)**

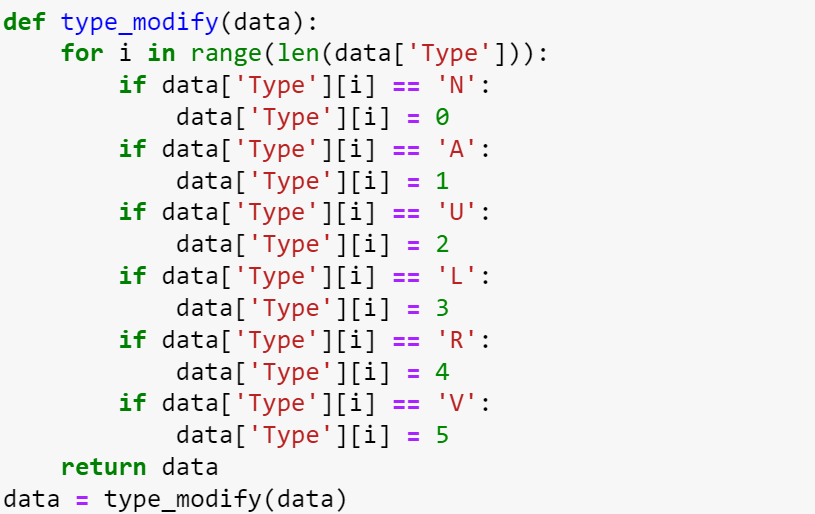


There are some differences between these plots. Type N is relatively flat and can reach a maximum of about 1200 in places with large fluctuations. Type L is not as smooth as N, with larger fluctuations during rest periods and increasing fluctuations. Type R is a bit similar to Type N, but its peak value is only around 1000, and the difference between peak and valley values is relatively large. Types A and R are similar, but A is more volatile and noisier than R. Type V is the most irregular of all types, while Type U has two peaks and is easier to identify.

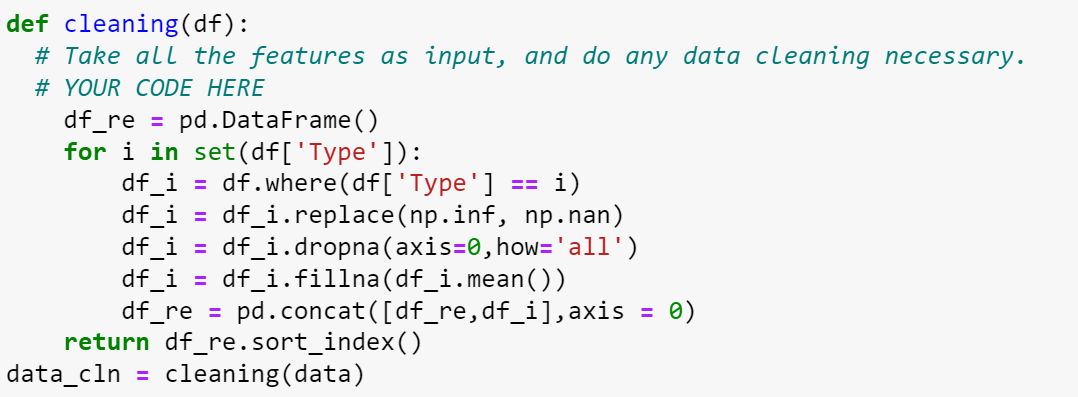
And the total number of data is 97838.



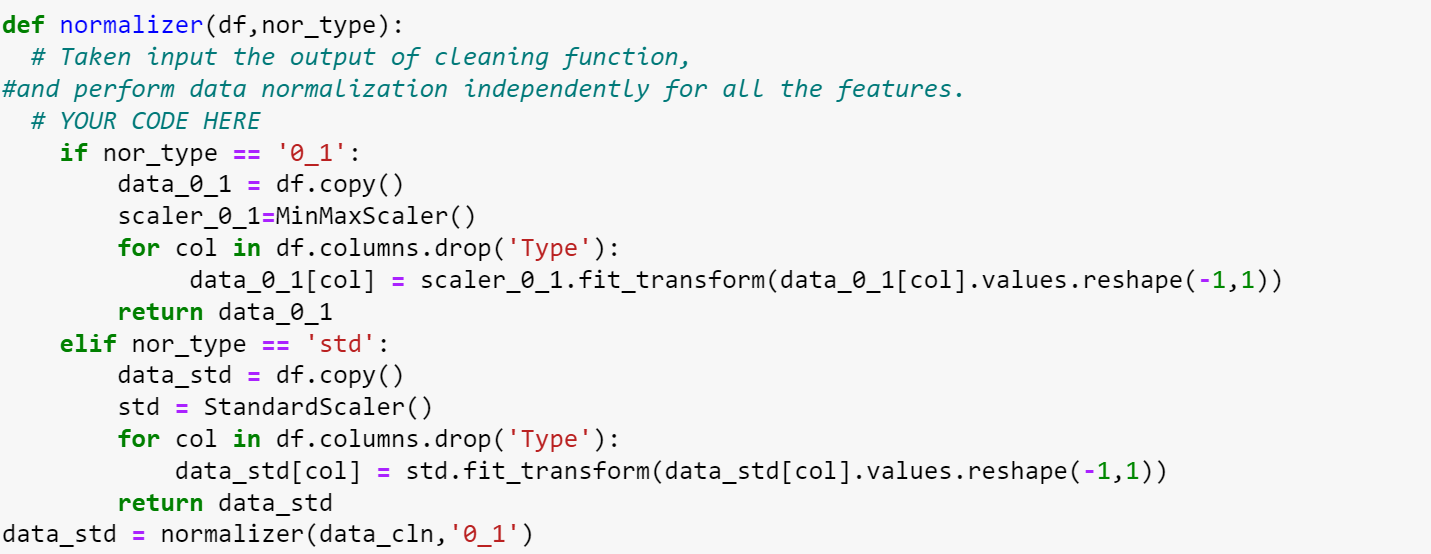
**Answer2)**



I transferred the target type from letters to numbers.

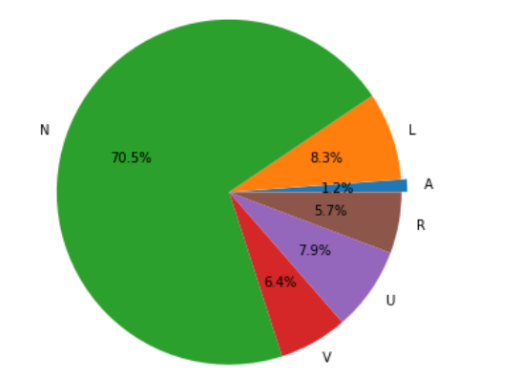
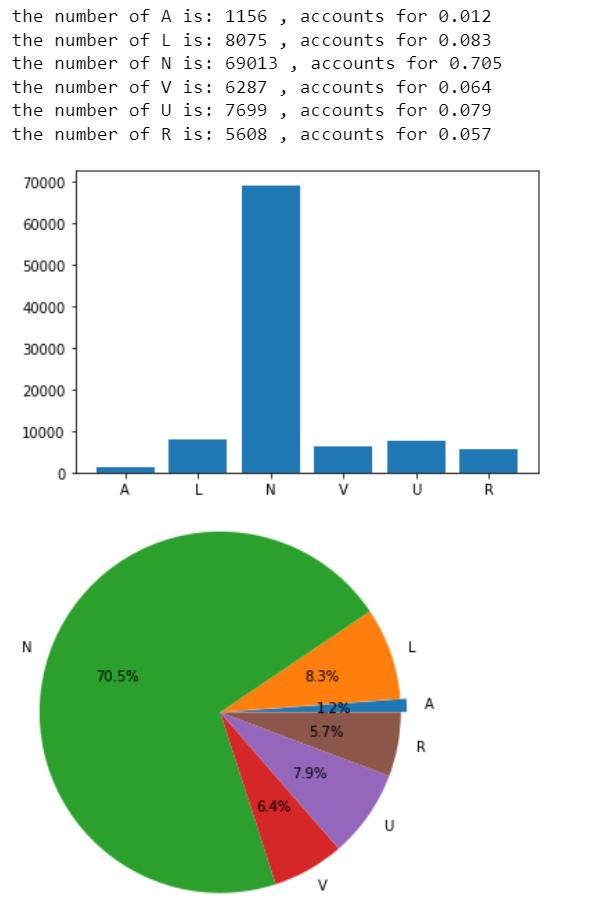


Cleaned the data by removing infinity values and none values.



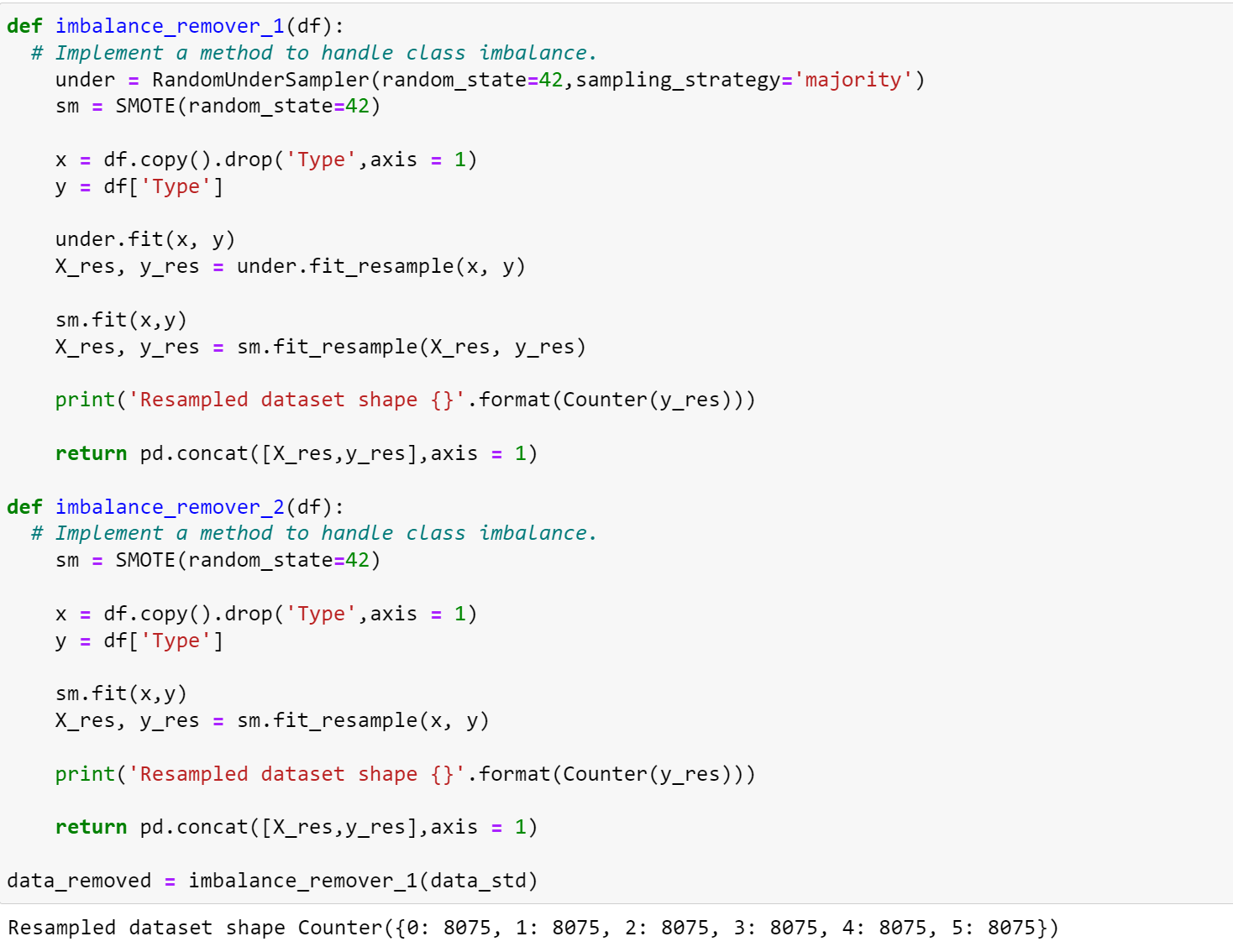
Then used MinMaxScaler() or StandardScaler() to normalize the data

**Answer3)**



The dataset is imbalanced. N is the largest type which accounts for over 70%. So an imbalanced remover is needed in this case.

Then two imbalanced\_remover are implemented.

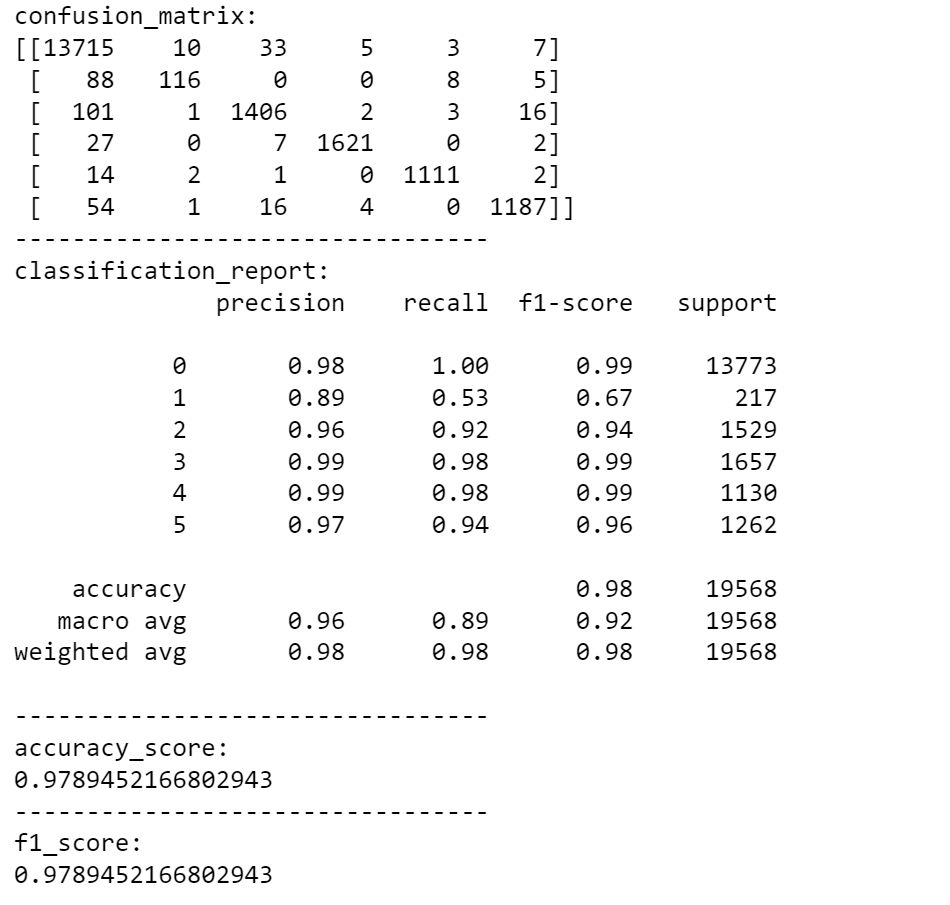


The dataset changed to 8075 for each type.

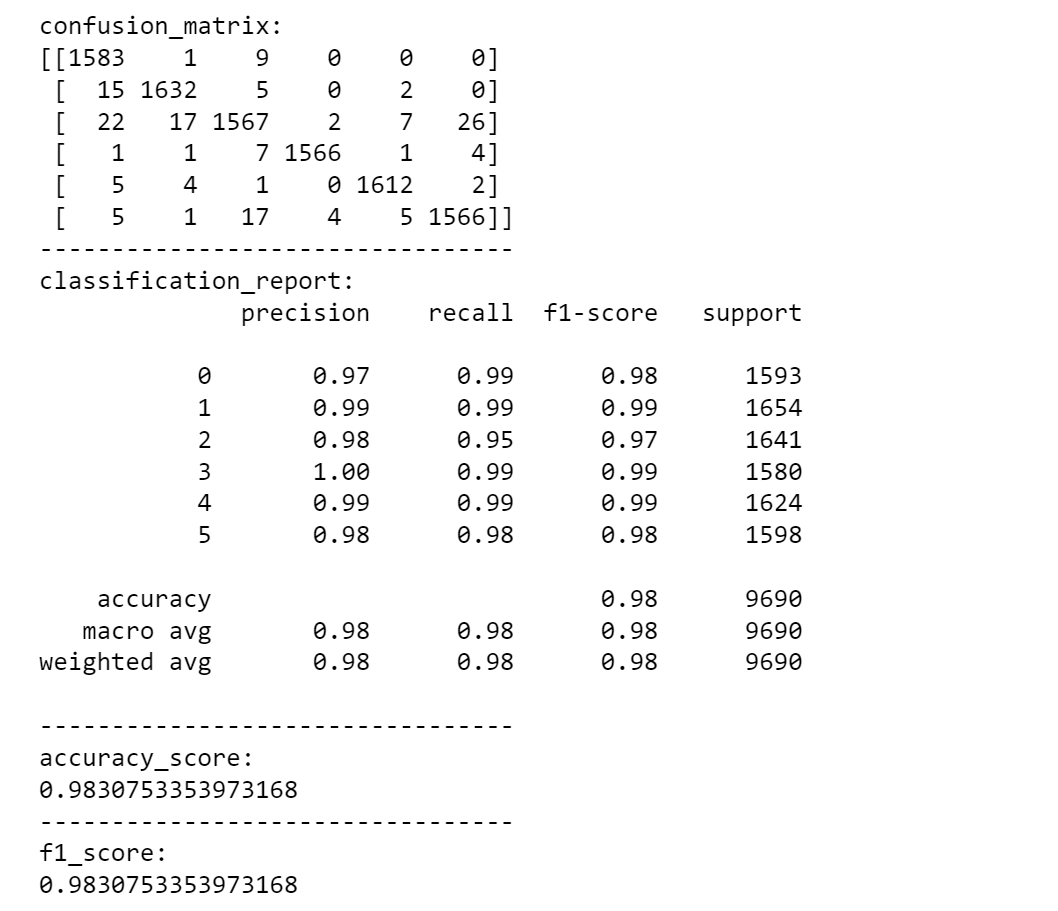
And the AutoEnconder.



Before the data augment, the metrics of the test set are as follows:



After that, they are:



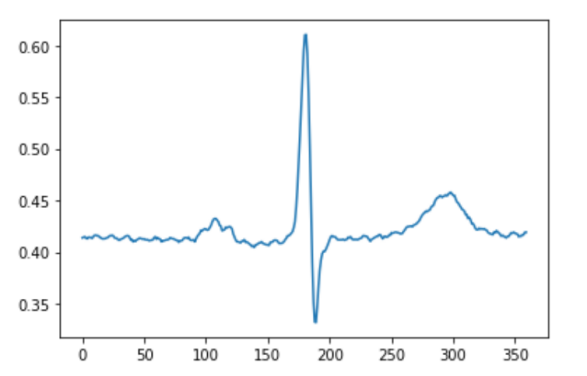
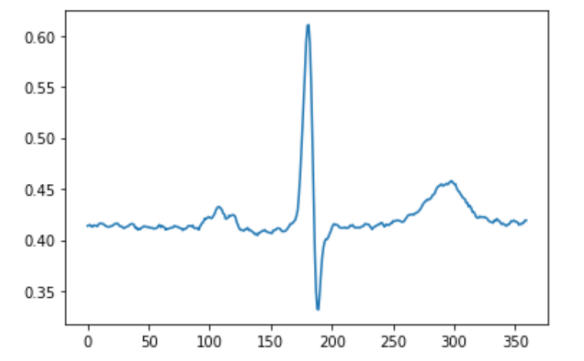
So I suppose that they work well.

**Answer4)**



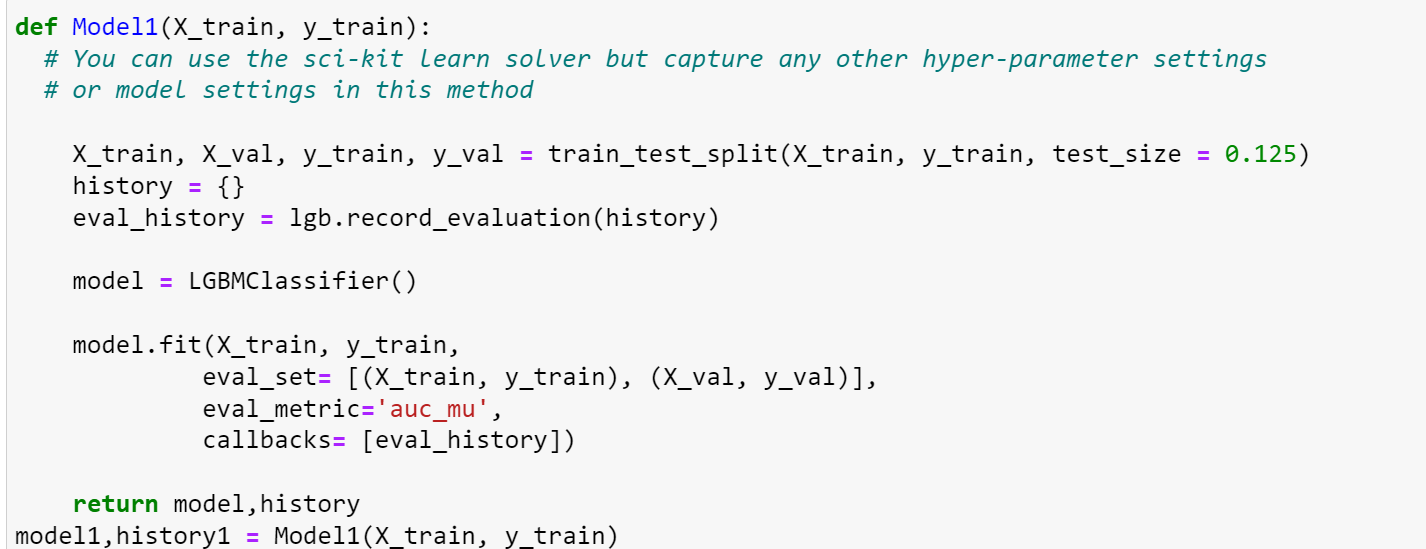
I used WT, named Wavelet transform, which is a new transform analysis method. It inherits and develops the idea of localization of short-time Fourier transform, and at the same time overcomes the shortcomings of window size that does not change with frequency, and can provide a "time-frequency" window that changes with frequency. , is an ideal tool for signal time-frequency analysis and processing. Its main feature is that it can fully highlight the characteristics of certain aspects of the problem through transformation, can analyze the localization of time (space) frequencies, and gradually refine the signal (function) through scaling and translation operations. Time subdivision, frequency subdivision at low frequency, can automatically adapt to the requirements of time-frequency signal analysis so that it can focus on any details of the signal.

Before and after denoising:

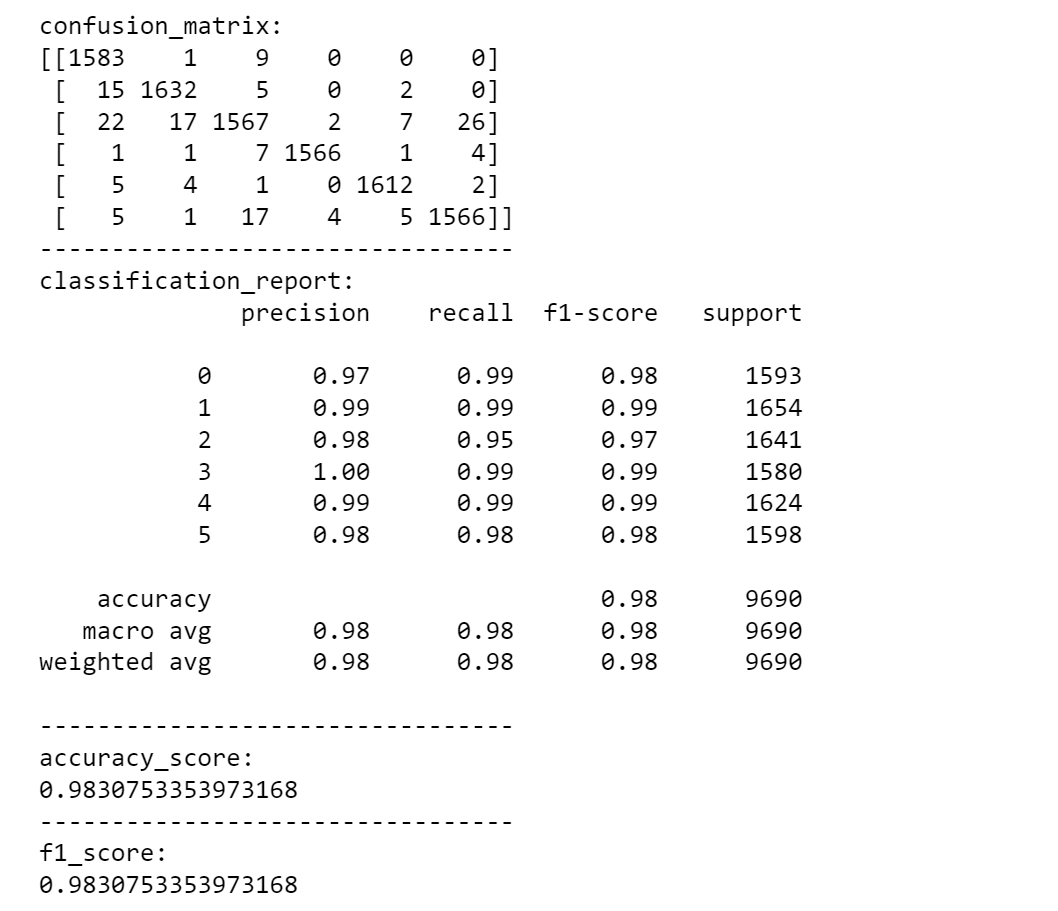


As we can see from the picture, The second one is more clear, and some noise has been removed, making the signal smoother, and there is not too much noise-causing over-fitting.

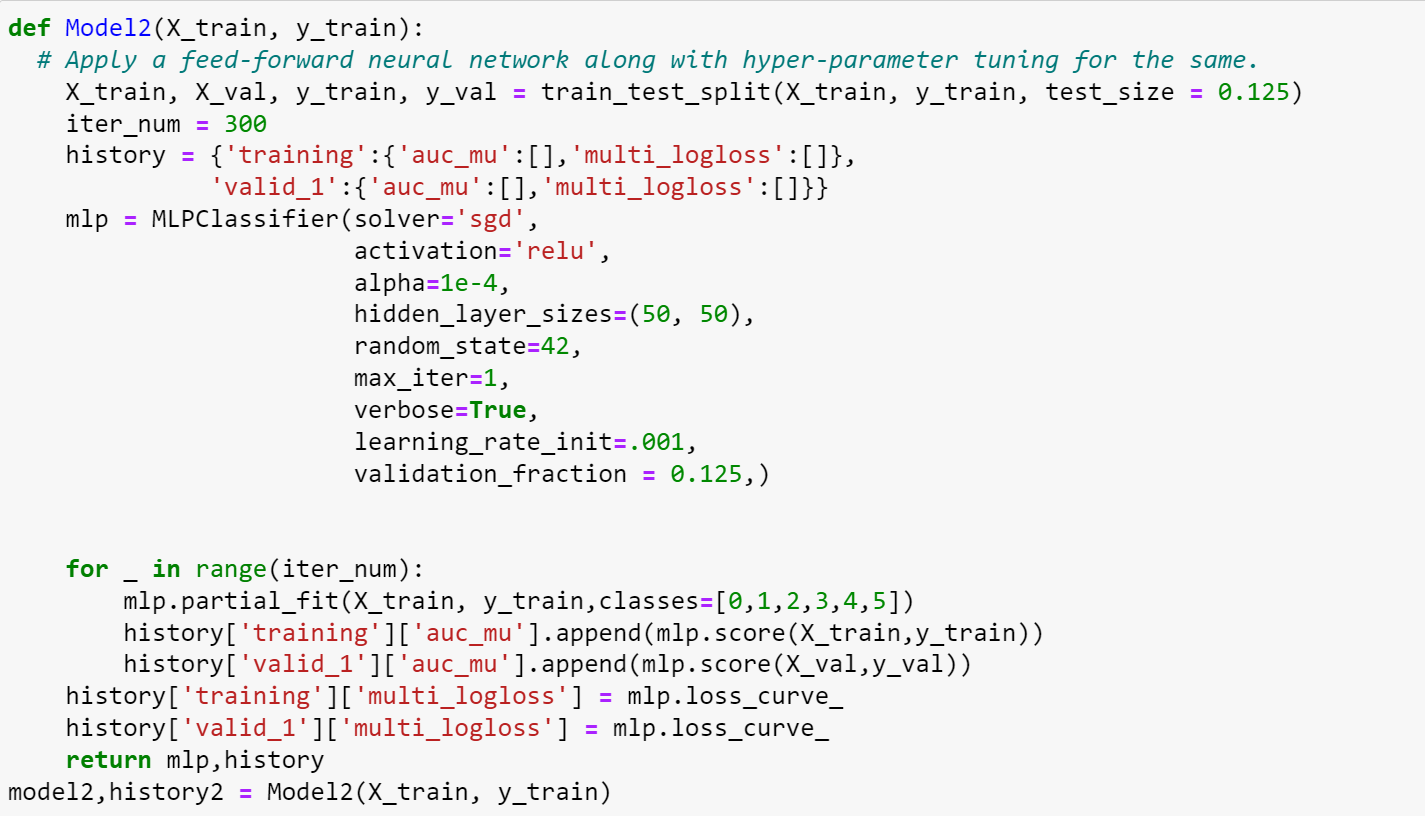
**Answer5)**



I chose lightBGM to train and got awesome precision on the test set, which is over 0.98. As for the parameters, just using its default values is good.

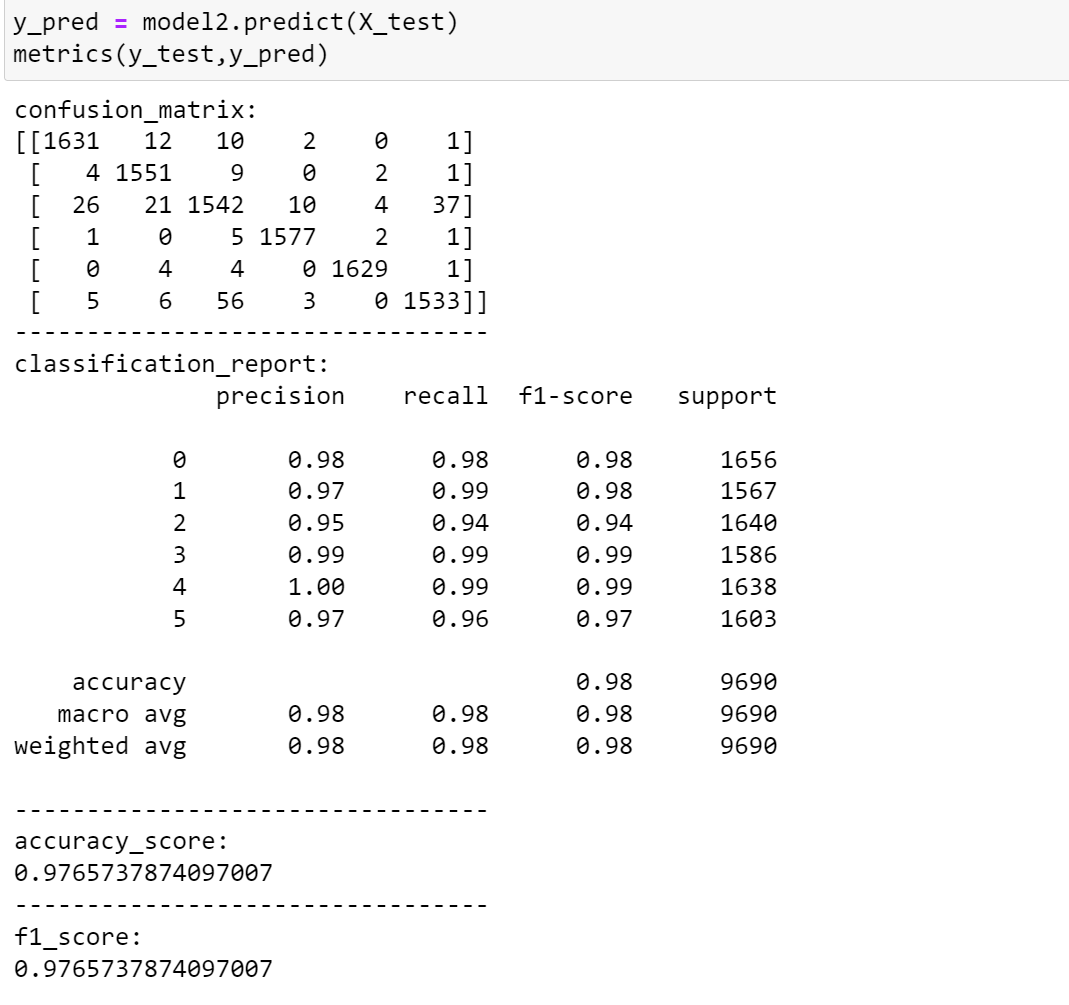


**Answer6)**



I used an MLPclassifier for this question and tuned its hyper-parameters, with alpha being 0.0001, and learning\_rate\_init being 0.001.

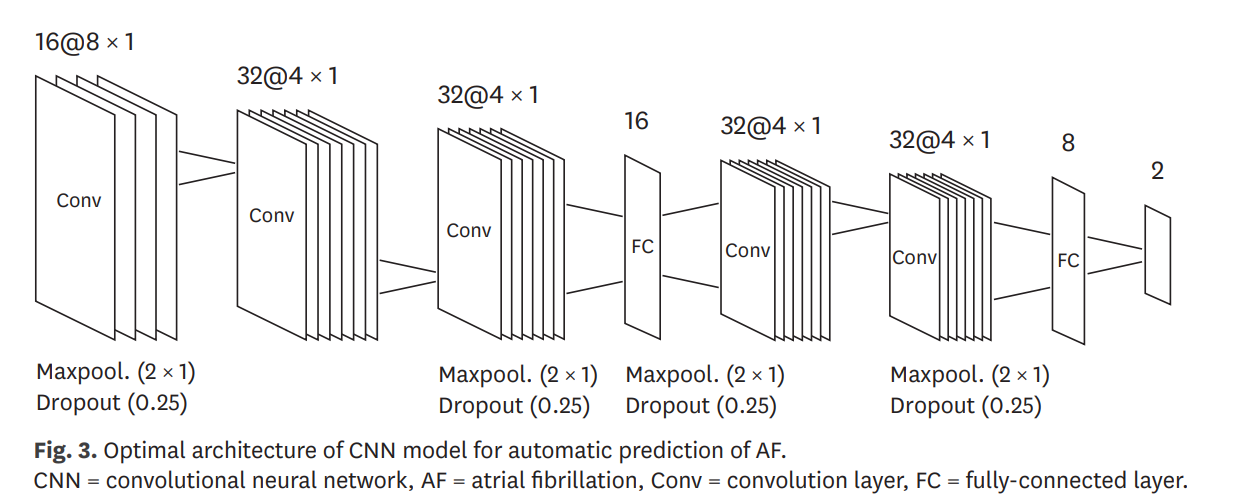
And the result is as follows:



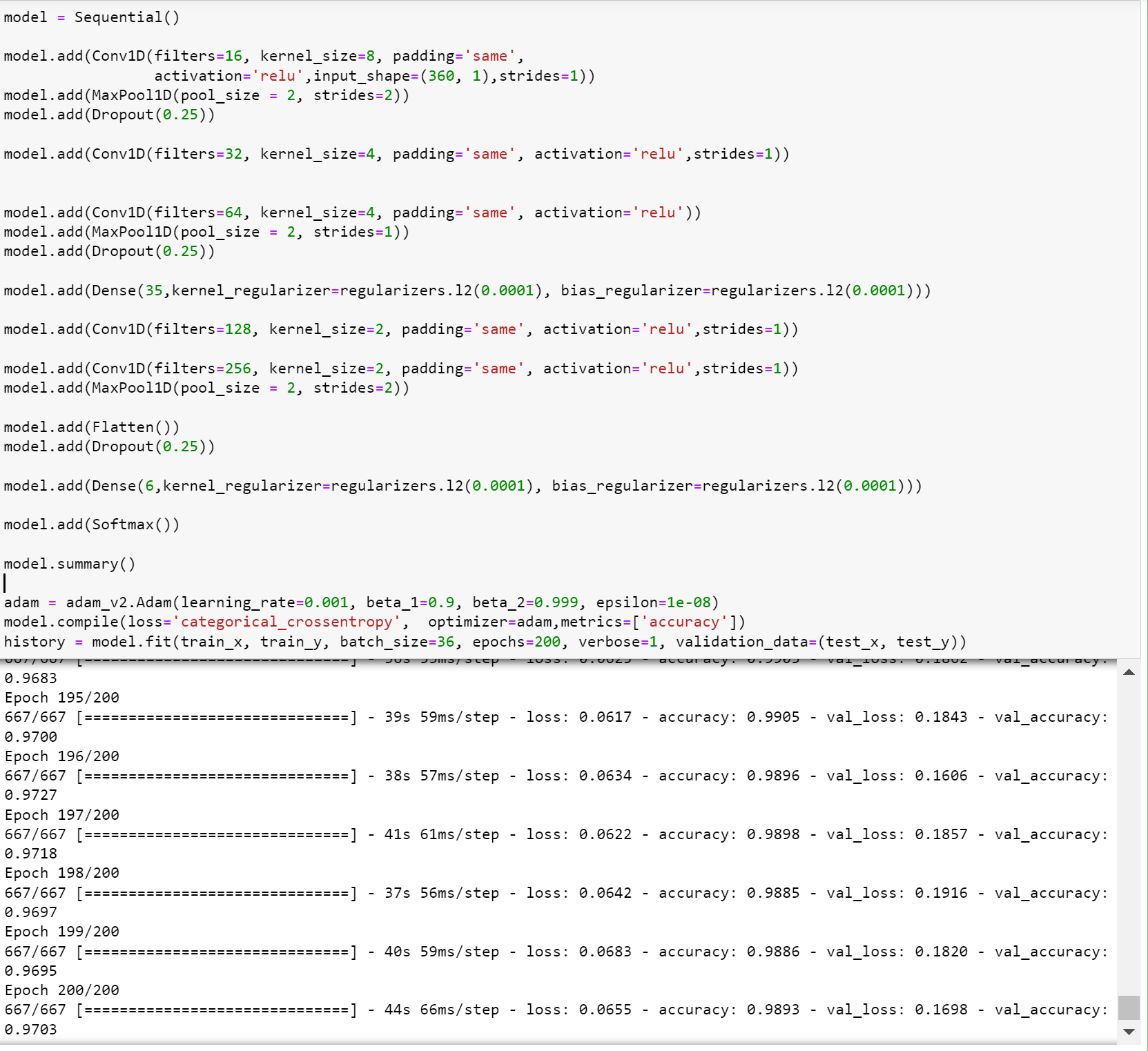
It works well.

**Answer7)**

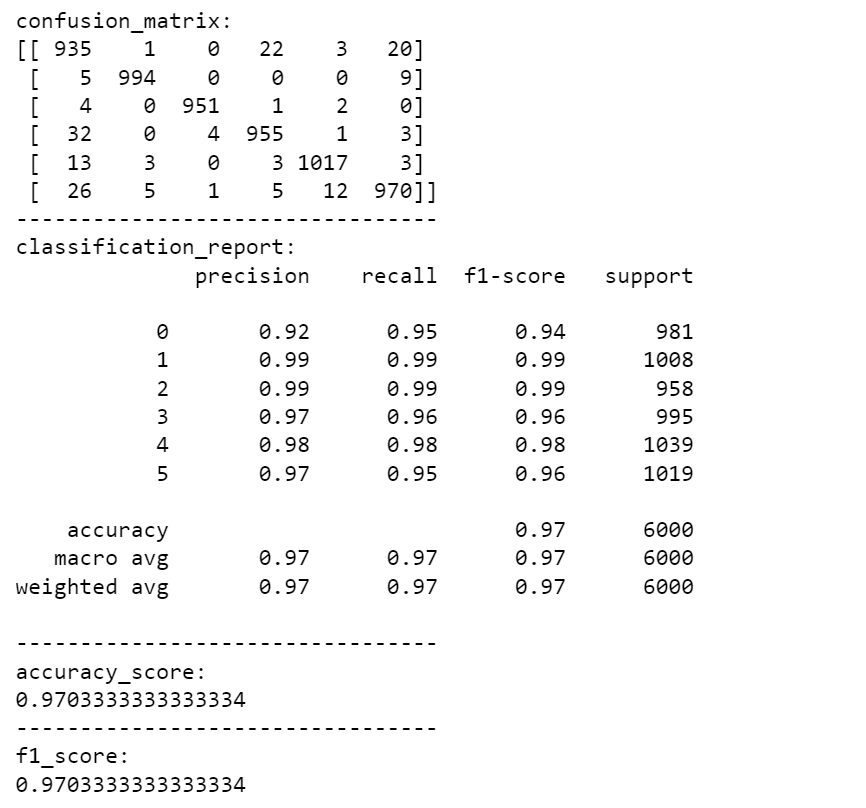
The paper[1] I used is called *Automatic Prediction of Atrial Fibrillation Based on Convolutional Neural Network Using a Short-term Normal Electrocardiogram Signal,* which implemented a convolutional neural network like the following graph:



And I tried to re-build it:



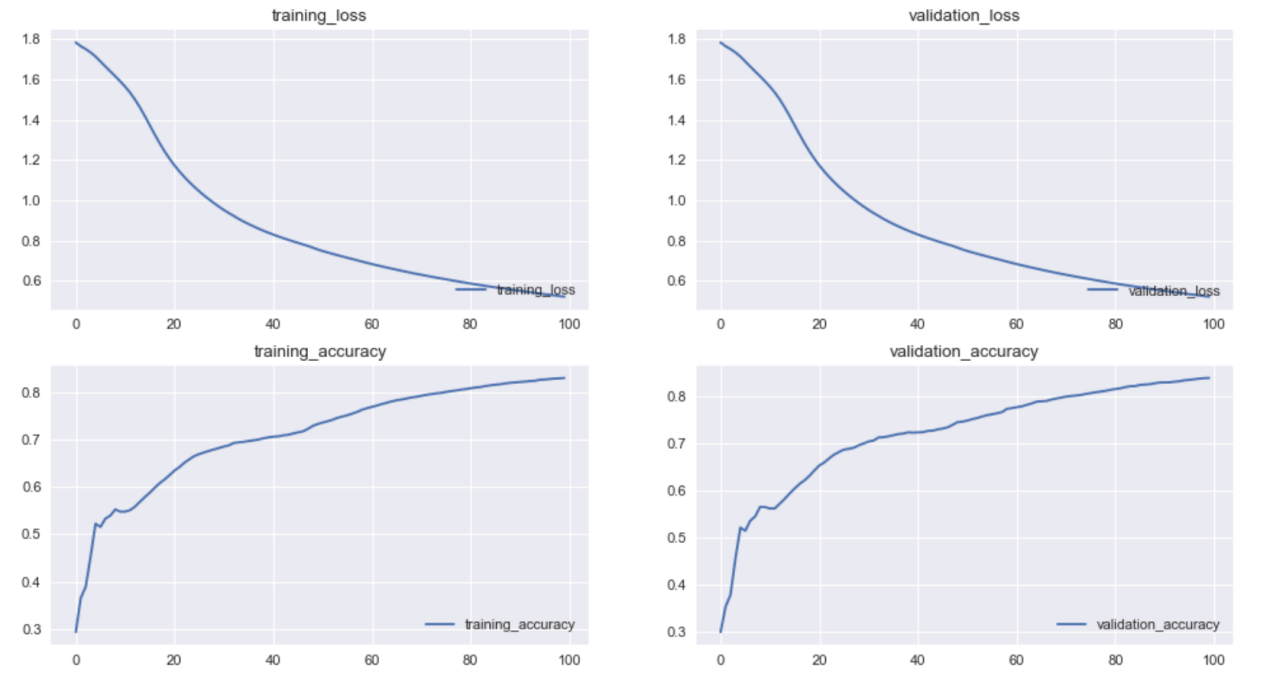
The matrics of the test set are:



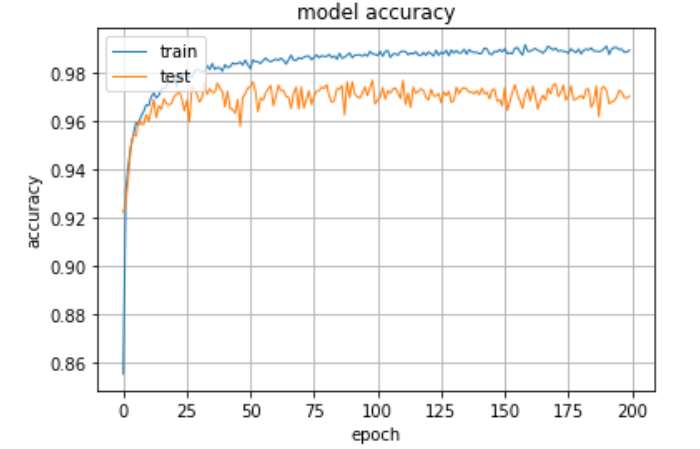
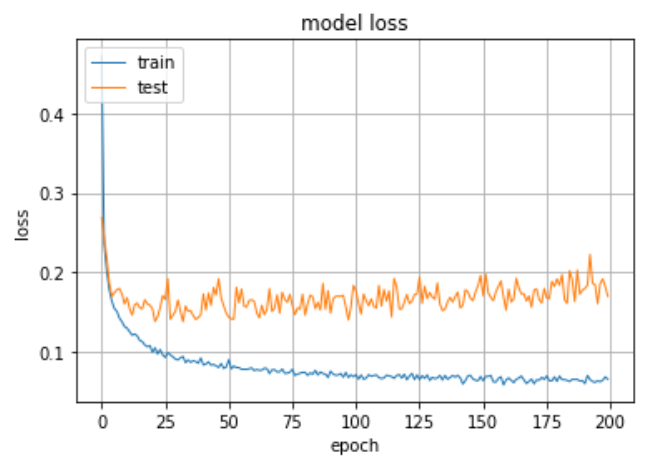
I got 0.97 which is lower than the accuracy of 98.6% in the parer. I suppose that my pre-processing methods are different than the methods referred to in the paper so my accuracy is still needed to be improved further.

**Answer8)**

MLPclassifier:



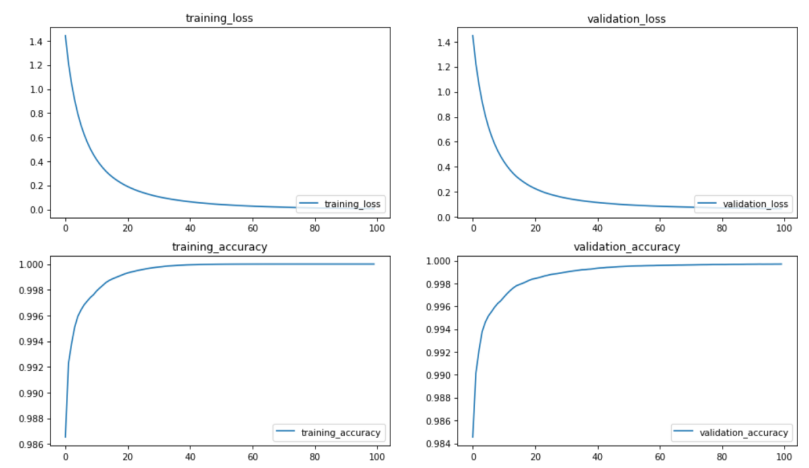
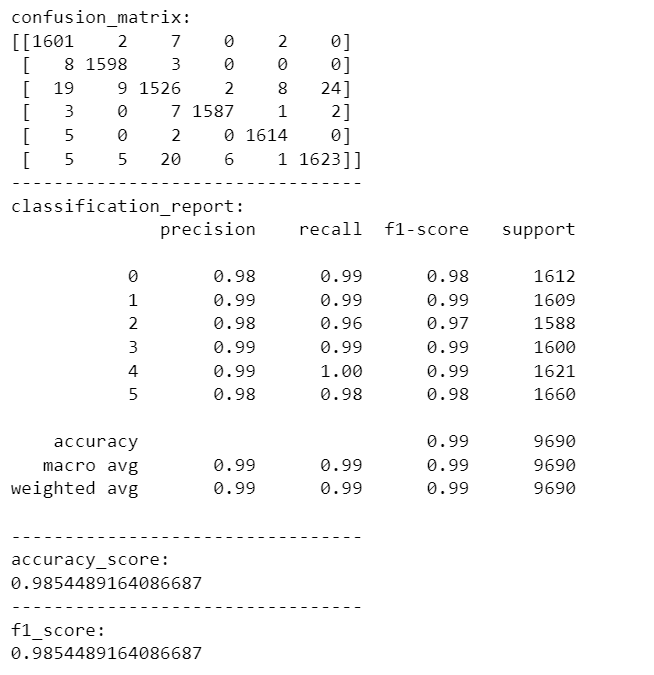
CNN in the paper:



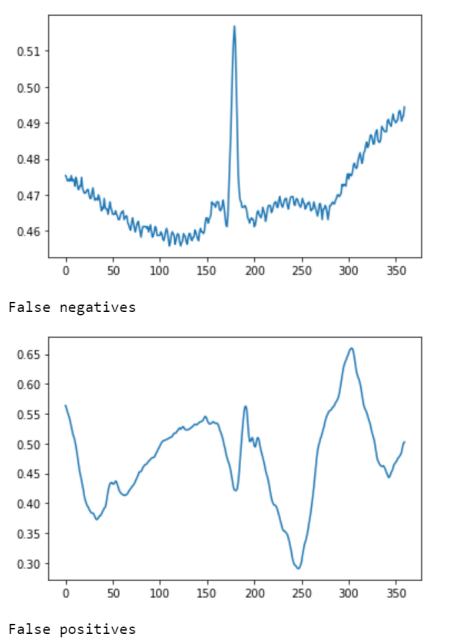
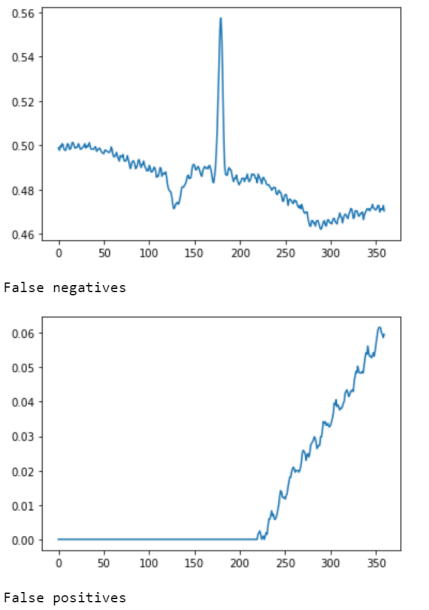
**Answer9)**

I chose V1 since I guess it is relevant to the heartbeat waves. And the V1 indicator has the largest number among other indicators. I matched V1 to those who have the value of V1 and then filled others with Nan. So for now it has 720 features.

Here is the result of training through the new dataset, it performs better than before.



**Answer10)**



False-negative: similarities with negatives, such as abnormal trends and huge fluctuations, are easily identified as false negatives.

False-positive: I think it may be because the amount of positive data is not complete, the machine cannot fully recognize it, and some fake data is created during data enhancement so that the machine cannot correctly identify the true positive.

**Teamwork**

Our team has two members: XiangyuGao and RuiLi. We worked together and talked about the assignment. In each part of this project, we did the tasks together. All the codes are programmed after we brainstorm, and there is no division of labor. For example, you do the first part and I do the second part. So we are very good at grasping the whole. Through this mission, we got to know each other better and became more familiar with each other, which provided an emotional foundation for future study and life, and I believe that we will work together and make progress together.

**References**

[1]Erdenebayar U, Kim H, Park JU, Kang D, Lee KJ.   Automatic Prediction of Atrial Fibrillation Based on Convolutional Neural Network Using a Short-term Normal Electrocardiogram Signal.   J Korean Med Sci. 2019 Feb;34(7):e64.   [https://doi.org/10.3346/jkms.2019.34.e64](https://doi.org/10.3346/jkms.2019.34.e64" \t "https://jkms.org/_blank)